The *Situate AI* Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder Early-stage Deliberations Around Public Sector AI Proposals

Anna Kawakami
Carnegie Mellon University
Pittsburgh, PA, USA
akawakam@andrew.cmu.edu

Amanda Coston
Microsoft Research
Cambridge, MA, USA
amandacoston@microsoft.com

Haiyi Zhu*
Carnegie Mellon University
Pittsburgh, PA, USA
haiyiz@cs.cmu.edu

Hoda Heidari*
Carnegie Mellon University
Pittsburgh, PA, USA
hheidari@cmu.edu

Kenneth Holstein*
Carnegie Mellon University
Pittsburgh, PA, USA
kjholste@andrew.cmu.edu

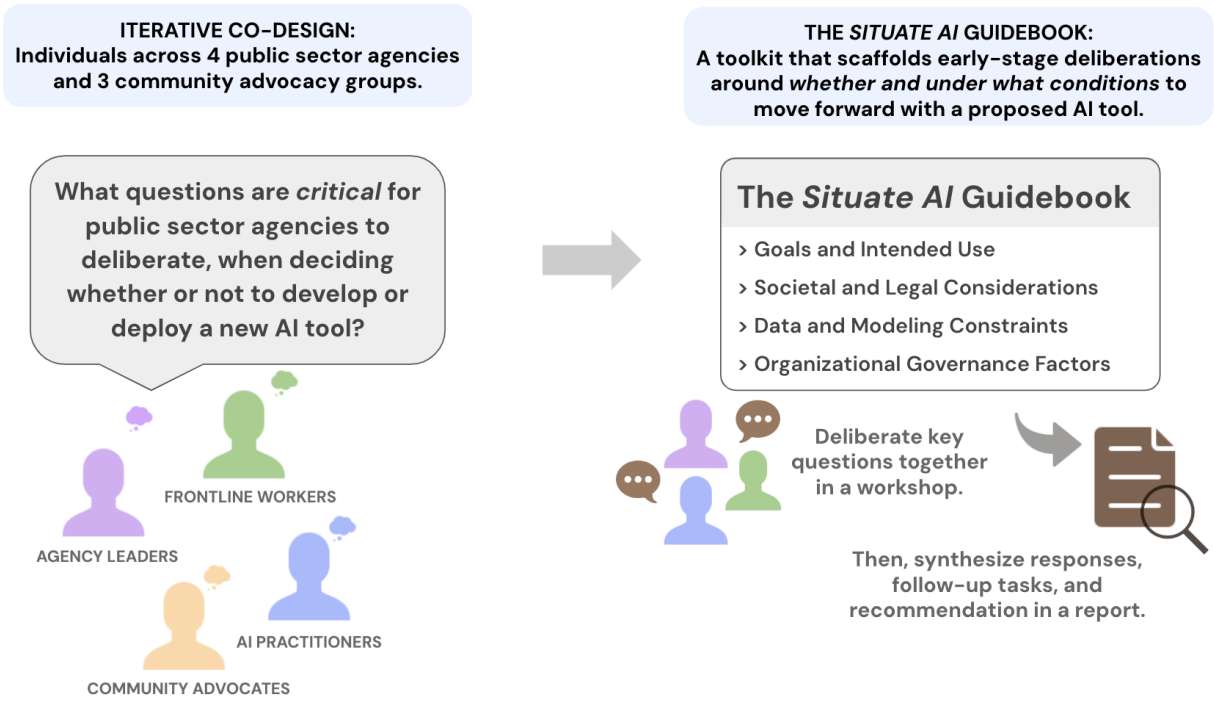

Figure 1: We conducted co-design activities and semi-structured interviews with public sector agency workers (agency leaders, AI practitioners, frontline workers) and community advocates to understand the questions they believed were *critical to discuss* yet currently overlooked before deciding to move forward with a public sector AI proposal. The *Situate AI* Guidebook synthesizes these key considerations into a toolkit to scaffold early-stage deliberations around *whether* and *under what conditions* to move forward with developing or deploying a proposed public sector AI tool.

*Co-senior authors contributed equally to this research.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642849

ABSTRACT

Public sector agencies are rapidly deploying AI systems to augment or automate critical decisions in real-world contexts like child welfare, criminal justice, and public health. A growing body of work documents how these AI systems often fail to improve services in practice. These failures can often be traced to decisions made during the early stages of AI ideation and design, such as problem formulation. However, today, we lack systematic processes to

support effective, early-stage decision-making about *whether* and *under what conditions* to move forward with a proposed AI project. To understand how to scaffold such processes in real-world settings, we worked with public sector agency leaders, AI developers, frontline workers, and community advocates across four public sector agencies and three community advocacy groups in the United States. Through an iterative co-design process, we created the *Situate AI* Guidebook: a structured process centered around a set of deliberation questions to scaffold conversations around (1) *goals and intended use* for a proposed AI system, (2) *societal and legal considerations*, (3) *data and modeling constraints*, and (4) *organizational governance factors*. We discuss how the guidebook's design is informed by participants' challenges, needs, and desires for improved deliberation processes. We further elaborate on implications for designing responsible AI toolkits in collaboration with public sector agency stakeholders and opportunities for future work to expand upon the guidebook. This design approach can be more broadly adopted to support the co-creation of responsible AI toolkits that scaffold key decision-making processes surrounding the use of AI in the public sector and beyond.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Public Sector AI, Participatory Approaches to Design, Responsible AI, Technology Governance and Policy

ACM Reference Format:
Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. 2024. The *Situate AI* Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder Early-stage Deliberations Around Public Sector AI Proposals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3613904.3642849

1 INTRODUCTION

Public sector agencies in the United States are rapidly adopting AI systems to assist or automate services in settings such as child welfare, credit lending, housing allocation, and public health. These tools have been introduced to help overcome resource constraints and limitations in human decision-making [12, 36]. However, as a growing body of work has documented, public sector AI tools have often failed to produce value in practice, instead exacerbating existing problems or introducing new ones [10, 30, 42, 65]. For example, the Michigan Unemployment Insurance Agency developed an AI-based fraud detection system (MiDAS); the agency stopped using the tool after realizing it falsely flagged over 90% of its cases—a discovery that was made only *after* the tool had been in use for over two years, impacting hundreds of thousands of people along the way [6]. Similarly, following the deployment of an AI-based tool for child maltreatment screening, Allegheny County's Department of Human Services faced significant criticism after the tool was found to exacerbate biases against Black and disabled communities [17, 20, 24, 25]. Research indicates that these problems in deployment were a consequence of fundamental conflicts between the tool's design on the one hand, and data limitations and worker needs on the other [17, 20, 30, 31]. Many other public sector agencies have dropped deployed AI tools for similar reasons, even after investing significant resources into their development (e.g., [28, 57]).

Many failures in public sector AI projects can be traced back to decisions made during the earliest problem formulation and ideation stages of AI design [13, 47, 65, 68]. AI design concepts that make it to production may be "doomed to fail" from the very beginning, for a variety of reasons. For example, AI design concepts have often been conceived in isolation from workers' actual decision-making tasks and challenges, leading to AI deployments that are not actually viable in practice [26, 31, 62, 67, 68]. Similarly, teams often propose design concepts for new tools that cannot possibly be implemented in an effective, safe, or valid way given technical constraints, such as the availability and quality of data [13, 50, 65, 68]. However, discussion of such constraints is commonly left to later stages of the AI lifecycle, by which point teams have invested in an idea and may be more reluctant to explore alternative ideas [30, 68]. While agencies utilizing AI may be motivated to try to mitigate issues at later project stages, such attempts are unlikely to yield meaningful improvements if fundamental issues around the problem formulation and solution design are left unaddressed [20, 31, 61, 65, 68].

In this paper, we ask: **How can we support public sector agencies in deciding whether or not a proposed AI tool should be developed and deployed in the first place?** Today, we lack systematic processes to help agencies make informed choices about which AI project ideas to pursue, and which are best avoided. As AI tools proliferate in the public sector, the failures discussed above indicate that agencies are repeatedly missing the mark with AI innovation. While existing responsible AI toolkits have provided guidance on ways to support AI development and implementation to ensure compliance with the relevant principles and values (e.g., [18, 37, 39, 53]), most existing toolkits are designed for use in industry contexts. Furthemore, most toolkits start from the assumption that the decision to develop a particular AI tool has *already* been made.

To address these gaps, we introduce the *Situate AI* Guidebook: a toolkit to scaffold early-stage deliberations around *whether and under what conditions* to move forward with the development or deployment of a proposed public sector AI innovation. To ensure that our guidebook and process design is informed by existing organizational needs, practices, and constraints in the public sector, we partnered with 32 individuals, spanning a wide range of roles, across four public sector agencies and three community advocacy groups across the United States. Over the course of 8 months, we iteratively designed and validated the guidebook with a range of stakeholders, including (1) public sector agency leadership, (2) AI developers, (3) frontline workers, and (4) community advocates. The public sector agencies we partnered with represent different levels of experience and maturity with AI development and deployment: At the time of this research, some had just begun ideating ways to integrate AI tools into their agencies' processes; some were already in the process of developing new AI tools; and some had already experienced failures in AI tool deployment that led to halts in their use. The community advocacy groups include organizations that, among other areas of focus, represent and support community

members in navigating challenging interactions with public services (e.g., parents negatively impacted by the child welfare system).

We conducted formative semi-structured interviews and iterative co-design activities that guided the content and process design of the *Situate AI* Guidebook. In particular:

- Through semi-structured interviews, we developed an understanding of public sector agencies' current practices and challenges around the design, development, and evaluation of new AI tools, in order to identify opportunities for new processes to improve current practice.
- Through co-design activities, participants ideated and iterated upon a set of questions that they believed were critical to consider before deciding to move forward with the development of a proposed AI tool. In addition, they described how they envisioned a deliberation process could be effectively structured for adoption at their agencies.

The resulting set of deliberation questions spanned a broad range of topics, from centering community needs to surfacing potential agency biases, given their positionality—topics which are relatively understated in existing Responsible AI toolkits developed for industry contexts. Notably, participants gravitated toward deliberation questions that promoted reflection on potential differences in perspective among the various stakeholders of public sector agencies (e.g., agency workers, frontline workers, impacted community members), surrounding topics such as the problem to be solved by an AI tool, notions of "community", or understandings of what it means for decision-making to be "fair" in a given context. This work presents the following contributions:

(1) **The *Situate AI* Guidebook**[1] **(*Ver*.1.0)** – the first toolkit co-designed with public sector agencies and community advocacy groups to scaffold *early-stage deliberations* regarding *whether or not* to move forward with the development of a proposed AI tool.
(2) **A set of 132 co-designed deliberation questions** spanning four high level topics, (1) *goals and intended use*, (2) *societal and legal considerations*, (3) *data and modeling constraints*, and (4) *organizational governance factors*. Participants indicated these considerations are *critical to discuss* when deciding to move forward with with the development of a proposed AI tool, yet are not proactively or deliberately discussed today.
(3) **Guidance on the overall decision-making process** that the *Situate AI* Guidebook can be used to support, informed by how participants envisioned they would use the guidebook in their agencies and by prior literature discussing related challenges that threaten the practical utility of research-created artifacts [37, 66].
(4) **Success criteria for using the guidebook** informed by participants' existing challenges, prior literature, and signals that participants themselves described as valuable in assessments regarding the guidebook's ability to promote meaningful improvements in their agency.

In the following sections, we first overview relevant bodies of prior literature to help ground and motivate the creation of our toolkit (Section 2). We then describe the approach we took to collaboratively develop the *Situate AI* Guidebook (Section 3), and describe the guidebook's major components, including its guiding design principles (Section 4.1), deliberation questions (Section 4.2), process design (Section 4.3), and success criteria (Section 4.4). We conclude with a discussion of anticipated challenges, as well as directions for future research aimed at understanding how to implement such deliberation processes most effectively. We also discuss implications for future co-design of responsible AI toolkits intended to promote meaningful change in public sector contexts (Section 5). The public sector agencies we partnered with in this study plan to explore the use of the guidebook through pilots, to identify further avenues for improvement.

[1] https://annakawakami.github.io/situateAI-guidebook/

2 BACKGROUND

2.1 Public Sector AI and Overcoming AI Failures

In the United States, public sector agencies are government-owned or affiliated organizations occupying the federal, state, county, or city government, responsible for making decisions around the allocation of educational, welfare, health, and other services to the community [41]. Public sector agencies across the United States are exploring how to reap the benefits of AI innovations for their own workplaces. AI tools promise new opportunities to improve the efficiency of public sector services, for example, by increasing decision quality and reducing agency costs [7, 11, 46, 63]. In 2018, 83% of agency leaders indicated they were willing or able to adopt new AI tools into their agency [5]. In the public sector, there is also a recognition that developing AI tools in-house can help ensure that they are better tailored to meet agency-specific needs, ensure they are trained on representative datasets, and account for local compliance requirements [16]. However, achieving responsible AI design in the public sector has proven to be an immense challenge [22, 23, 56, 64]. The domains where agencies are attempting to apply AI are often highly socially complex and high-stakes–including tasks like screening child maltreatment reports [58], allocating housing to unhoused people [35], predicting criminal activity [32], or prioritizing medical care for patients [45]. In these domains, where some public sector agencies have a fraught history of interactions with marginalized communities [4, 54], it has proven to be particularly challenging to design AI systems that avoid further perpetuating social biases [10], obfuscating how decisions are made [31], or relying on inappropriate quantitative notions of what it means to make accurate decisions [13]. Public sector agencies are increasingly under fire for implementing AI tools that fail to bring value to the communities they serve, contributing to a common trend: AI tools are implemented then discarded after failing in practice [21, 56, 67, 68].

Research communities across disciplines (e.g., HCI, machine learning, social sciences, STS) are beginning to converge toward the same conclusion: **Challenges observed downstream can be traced back to decisions made during early problem formulation stages of AI design**. Today, we lack concrete guidance to support these early stages of AI design [13, 47, 65, 68]. For example, after observing decades of failures to develop clinical decision support tools that bring value to clinicians, researchers have found that AI developers may **lack an adequate understanding of**

which tasks clinicians desire support for, leading to the creation of tools that target problems that clinicians do not actually have [15, 21, 67]. These trends are beginning to surface across other domains that have more recently begun to explore the use of new AI tools. In social work, researchers have found that technical design decisions in decision support tools reflect **misunderstandings around the type of work that social workers actually do**, leading to deployments where, for instance, the underlying logic of the model conflicts with how workers are trained and required to make decisions [30, 31]. Others have surfaced how seemingly technical design decisions made during early stages of model design actually embed **policy decisions that conflict with community values and needs** [20, 59, 61].

In addition to concerns regarding how well the problem formulation and design of a given AI tool reflects worker practices and community values, there is a concern that AI tools deployed in complex, real-world domains may be **conceived without adequate consideration for the actual capabilities of AI**. For example, examining a range of real-world decision support tools (e.g., in criminal justice, child welfare, tax lending), researchers have argued that existing AI deployments lack validity, due to limitations in the types of data that can be feasibly collected to train the desired model [13, 42, 50, 65]. This highlights the need for developers and organizations to reflect upon technical constraints and limitations at earlier stages of the AI development lifecycle, such as when evaluating whether or not to pursue a proposed AI project in the first place.

While public sector agencies have emphasized the potential to improve decision accuracy and reduce bias as a key motivation to use new AI tools (e.g., [14]), these challenges around the problem formulation and design of AI systems implicate the veracity of these claims [13, 31, 50, 61, 65, 68]. We identify a significant opportunity to better support public sector agencies in making systematic, deliberate decisions regarding whether or not to implement a given AI tool proposal. Given the vast potential for harm, and the similarly vast potential for AI systems to meaningfully support workers and improve services in the public sector, it is critical to support agencies through concrete guidance and processes in making more informed decisions around which AI tool proposals to pursue, and which to avoid.

2.2 Toolkits for Responsible AI Governance

In an effort to support responsible design and development of AI systems in practice, the HCI, ML and FAccT research communities have contributed a range of responsible AI toolkits. These toolkits are intended to support and document assessments of AI systems, including their (potential) impacts (e.g., [52]), intended use cases (e.g., [39]), capabilities and limitations (e.g., [53]), dataset quality (e.g., [19, 49]), and performance measures (e.g., [39]). Many of these toolkits are intended to be used as communication tools. For example, Model Cards provide a structure for communicating information regarding the intended uses, potential pitfalls, and evaluation measures of a given ML model, to support assessments of suitability for a given application and context of use [39]. Recent research surveying these toolkits have found that the majority of existing toolkits frame the work of AI ethics as "technical work for individual technical practitioners [66]. For the majority of existing responsible AI toolkits, the primary users are ML practitioners, limiting the forms of knowledge and perspectives that inform the work of "AI ethics'' [66]. A smaller number of toolkits have been designed for use by organization-external stakeholders, to support impacted end-users in interrogating and analyzing deployed automated decision systems (e.g., the Algorithmic Equity Toolkit [33]); provide impacted stakeholders with an opportunity to share feedback on an AI system's use cases and product design (e.g., Community Jury [1]); or support philanthropic organizations in vetting public sector AI technology proposals [3]). In all of these examples, the toolkit supports examinations of AI systems that have already been developed and sometimes even deployed.

Most existing toolkits assume that the decision to develop a particular AI system has already been made. Therefore, even when they are intended to support reflection and improvement of the AI system, the types of improvements that could stem from using the toolkit tend to be limited to those that would not require fundamental changes to the underlying technology. Meanwhile, while some existing responsible AI toolkits target earlier stages of AI development (e.g., [69]), these have primarily been designed for *private sector* contexts. Yet there is good reason to expect that public sector agencies would benefit from tailored responsible AI tools. For instance, compared with the private sector, there is a greater expectation that public sector agencies exist to serve people and are expected to make decisions that center communities' needs. When making decisions as critical as what new AI tools to deploy, agencies are expected to adhere strongly to values such as deliberative decision-making, public accountability, and transparency. To date, there exists minimal concrete and actionable guidance on how to support *public sector agencies* in scaffolding *early-stage* deliberation and decision-making.

A related existing artifact is the AI Impact Assessment, described as a "process for simultaneously documenting an [AI] undertaking, evaluating the impacts it might cause, and assigning responsibility for those impacts" [40]. AI Impact Assessments have been proposed for both public and private sector contexts, and are intended to be completed either at an early stage of AI design (e.g., [2]), or after an AI system is developed or deployed (e.g., [38]). Another related artifact is the Data Ethics Decision Aid (DEDA) [18], a framework to scaffold ethical considerations around data projects proposed in the Dutch Government. However, neither of these examples are designed as *deliberation toolkits*, to promote collaborative reflection and discussion around the underlying problem formulation or solution design of an AI tool. AI Impact Assessments and ethical decision aids have also not typically been designed in collaboration with the stakeholders they intend to serve. With recent research suggesting low adoption of responsible AI toolkits in real-world organizational contexts [51, 66], a co-design approach with organizational stakeholders has the potential to generate responsible AI tools that work in practice.

3 METHODS: CO-DESIGN AND VALIDATION OF THE *SITUATE AI* GUIDEBOOK

To iteratively co-design and validate the *Situate AI* Guidebook, we conducted semi-structured interviews and co-design activities with

a range of stakeholders both within and outside of public sector agencies. In this section, we describe participants' backgrounds, the approach and resources used in our iterative co-design process, and our data analysis approach.

3.1 Participants and Recruitment

We co-designed the *Situate AI* Guidebook with individuals from **four public sector agencies** across the United States. Collectively, this set of public sector agencies has experienced a range of decision-making scenarios around the creation or use of AI-based decision tools. All four agencies are currently ideating new forms of AI-based tools, three have already implemented AI tools, and at least one had previously deployed an AI tool and subsequently decided to abandon it. From these agencies, we wanted to include stakeholders at different levels of the organizational hierarchy including those with experience making relevant decisions and those who are involved in the development or consumption of AI tools but who are not typically involved in decisions around development and deployment. We therefore included participants from three core stakeholder groups: 1) **Agency leaders** (*L*) who are in director or managerial roles, typically involved in agency- or department-level decisions including whether to design and deploy a particular AI tool, 2) **AI developers, analysts, and researchers** (*A*) who are in development, analysis, or research teams internal to a given public sector agency and typically build and evaluate AI tools, and 3) **Frontline decision-makers** (*F*) whose occupations bring them in direct contact with the community their agency serves and whom an AI tool may be intended to assist. Because we wanted to learn from additional frontline decision-makers but had access only to a limited number at the public sector agencies we connected with, we recruited additional participants beyond these agencies, with relevant professional backgrounds. These included social work graduate students with prior field experience making frontline decisions in public sector agencies.

In addition, we co-designed the guidebook with individuals from **three community advocacy groups** across the United States, including family representation and child welfare advocacy groups. Individuals from these organizations created the fourth stakeholder group: 4) **Community advocates** (*C*) who represent and meet community members' needs around public services. While the Situate AI Guidebook is intended to be used by workers within a public sector agency, we included community advocates because we wanted the guidebook to represent their perspectives regarding the most critical considerations for moving forward with an AI tool design. As discussed in Section 4.3.2, we also worked with community advocates to begin envisioning what a future version of the toolkit, aimed at engaging community members in the deliberation process, might look like.

In total, 7 agency leaders; 7 developers, analysts, and researchers; 7 frontline decision-makers; and 11 community advocates participated in the co-design process. To recruit public sector agencies, we contacted 19 U.S. public sector agencies at the state, city, or county level with human services departments. We received responses from five agencies. Following a series of informal conversations to share our research goals and study plans, four of the agencies decided to participate in the study. To recruit individuals from community advocacy organizations, we contacted community leaders and advocates across 8 organizations. While we requested individual study participation, some participants preferred to participate in the research study in small groups. By participating in groups, they believed they could provide a more extensive set of insights together. 21 out of 25 sessions were conducted individually, and the remaining four were group interviews. For ease of communication, we will use the singular noun "participant" throughout the remainder of the paper.

3.2 Iterative Co-Design and Validation

The *Situate AI* Guidebook integrates findings across semi-structured interviews and co-design activities, which were conducted over the course of eight months between November 2022 and June 2023. The study sessions were ~90 minutes long for public sector workers, who were involved in both the interviews and co-design activities; the study sessions were ~60 minutes for community advocates, who were only involved in the co-design activities.

3.2.1 Formative Semi-structured Interviews. To ensure that the *Situate AI* Guidebook is designed to address real-world needs and goals, we conducted semi-structured interviews with public sector agency workers to understand (1) their existing challenges and barriers to making decisions around AI systems and (2) desires for improving their current decision processes. Specifically, to understand existing decision-making processes, we asked each participant to recall a specific prior experience in which they or their agency needed to decide whether to move forward with the development or use of a new AI-based tool. As participants shared their stories, we asked follow-up questions to probe on possible causes behind the challenges that they described. For example, after describing how they previously made a related decision, we asked "What's challenging to do well now, when you're making those decisions?'' or "What would you ideally want to discuss in conversations surrounding those decisions?'' If a participant shared that they had not personally been involved in decisions around AI design and deployment—as was the case with community advocates and many frontline workers—these questions would be skipped, and more time would be spent on discussing these participants' desires for improved decision processes. We report findings on how agency decision-makers currently make decisions around AI, including how their decisions are shaped by complex power relations they hold with stakeholders external and internal to their agency (e.g., legal systems, frontline workers), in [29]. In this paper, we share complementary findings that provide design rationale for the *Situate AI* Guidebook's design.

3.2.2 Co-Designing the Deliberation Questions. In the co-design activity, we first presented each participant with three potential scenarios: (1) Discussing *ideas* for new algorithms to improve services, (2) Deciding whether to pursue the *development* of a given algorithm design to improve services, and (3) Deciding whether to *adopt* an existing algorithm already implemented by others. We asked the participant to pick the scenario they had the most experience in or faced the most challenges for. For the scenario they

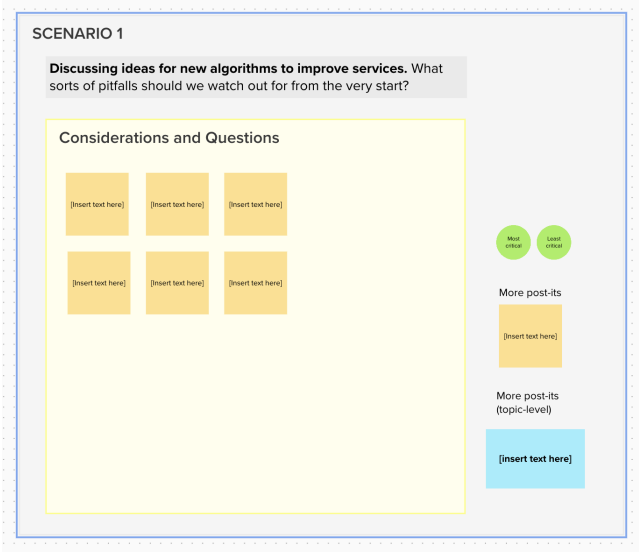

Figure 2: Screenshot of a blank board shown to participants at the start of the co-design activity on Mural. This board presents Scenario 1: Discussing ideas for new algorithms to improve services.

selected, we asked the participant to think about what critical considerations and questions they believe should be on the table, when deliberating around these scenarios in an ideal future situation. If the participant was having a challenging time thinking of potential considerations and questions, we provided them with examples that were directly based on challenges they had brought up during the semi-structured interview (if applicable). To help document and organize, in real-time, the considerations and questions the participant was bringing up, we shared our screen and a link to an online board on Mural, a collaborative web application where multiple users can generate and arrange sticky notes. See Figure 2 for an example of a blank canvas.

We asked each participant to brainstorm critical considerations and questions they would want future agencies to discuss. To avoid biasing participants, they were initially asked to openly ideate their own questions without viewing questions generated by prior participants. Following this, participants were shown existing questions, providing them an opportunity to comment upon and validate existing questions generated by other participants. As the participant openly generated ideas for questions, one of the members of our research team took post-it notes on what they were saying on the Mural board. The researcher would frequently check in with the participant, to ensure the post-its accurately represented their ideas. We also welcomed them to edit the post-its or create new ones. As they brainstormed, we asked follow-up questions to better understand how they think a given question could get answered, what makes it challenging to answer the question now, or how they are conceptualizing certain terms. For example, when a participant generated the question "How well are we involving community members in these decisions?," we asked them to further elaborate on what this might look like in practice. This generated additional post-its, like "How well do we understand the costs, risks, and effort required of community members, if we invite them to contribute to model design decisions?'' and "How are we weighting false positives and false negatives in a given algorithm, based on what type of mistake that is for the impacted community members?''

As mentioned above, after the participant generated their own questions and considerations on the blank canvas, they were shown a list of topics and example questions for additional consideration. This helped scaffold further ideation on any considerations they may have missed in their initial ideation. In the first study session, we provided an initial list of eight broad topics, informed by prior literature: (a) Overall goal for using algorithmic tool, (b) Selection of outcomes that the algorithmic tool should predict, (c) Empirical evaluations of algorithmic tool, (d) Legal and ethical considerations around use of algorithmic tool, (e) Selection of training data for algorithmic tool, (f) Selection of statistical models to fit data, (g) Long-run maintenance of algorithmic tool, and (h) Organizational policies and resources around use of algorithmic tool. We prompted the participant to discuss any new ideas the provided topical categories inspired, or any disagreements they had with the categories. Figure 3 shows an example of what this list looked like in later stages of the co-design process.

Between study sessions one or more researchers in our team iterated on the post-its generated during that study, to reduce redundancies and improve clarity. We then grouped the individual questions and considerations underneath the existing topical categories, while iteratively refining categories or creating new categories and subcategories as needed. The next participant was shown this updated version of the aggregated questions and topics at the end of the study.

3.2.3 Guidebook Reflection and Validation. Participants that contributed to later stages of the co-design process were shown an overview of the aggregated questions, a recommended deliverable for the deliberation guidebook, and a high-level outline of a proposed deliberation process. We first showed participants the aggregated questions, and asked if there were any questions that they felt were critical to include but missing. We additionally asked if they disagreed with the importance of any of the questions, or if the wording of any question was confusing in any way. We then showed the participant an overview of the deliberation process and asked for their perspectives around what they would like to change in the proposed process, to have it fit better into their existing organizational decision-making processes. To address challenges in the potential use of the protocol, we also asked participants (especially frontline workers and community advocates) about challenges they anticipate with participating in the deliberation process. We then invited participants to discuss potential adjustments to the process or alternative processes that can help address these challenges and create a safer environment for them.

3.3 Qualitative Analysis

The study recordings from the semi-structured interviews and co-design activities were transcribed and then qualitatively coded by two members of the research team using a reflexive thematic analysis approach [8]. We ensured that all interviews were coded by

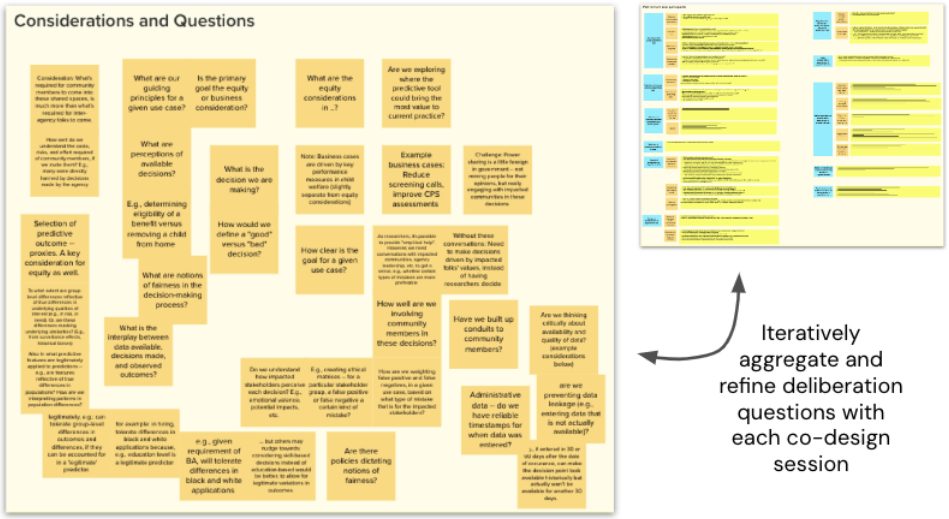

Figure 3: Screenshot of a Mural board populated with post-its after one participant's co-design activity. In our iterative co-design process, these post-its were refined by the research team then added to an aggregated list of questions that were successively grouped into higher level categories.

the first author, who conducted all of the interviews and, whenever applicable, another author who observed the interview. The first author coded one transcript first, then discussed the codes with other coders to align on coding granularity. Each coder prioritized coding underlying reasons why participants generated certain questions during the co-design activity, while also remaining open to capturing a broader range of potential findings. We resolved disagreements between coders through discussion.

drawing upon our thematic analysis. Where appropriate, we describe how the Situate AI Guidebook compares with existing responsible AI toolkits. At times, participants diverged in their desires (e.g., regarding how the decision-making process should be integrated into their agency). In some of these cases, our research team integrated these disagreements into the design of the guidebook (see Design Principle 2); in other cases, we document how these disagreements present challenges for the use of the guidebook, suggesting opportunities for future work (Section 4.3 and Section 5).

4 THE *SITUATE AI* GUIDEBOOK

The Situate AI Guidebook is a process to scaffold early-stage deliberations around *whether and under what conditions* to move forward with the development or deployment of a proposed AI innovation. The current version of this toolkit is intended for use within public sector agencies at various stages of maturity in their use of AI tools—from those that are just beginning to consider the use of new AI tools to those that may already have years of experience deploying AI tools. The deliberation questions are designed to be discussed across different stakeholders employed in a public sector agency, such as agency leadership, AI practitioners and analysts, program managers, and frontline workers.

In this section, we provide an overview of the Situate AI Guidebook as an outcome of our co-design and validation sessions. We describe the Situate AI Guidebook through the following sections: (Section 4.1) Guiding Design Principles, (Section 4.2) Content Design, (Section 4.3) Process Design, and (Section 4.4) Success Criteria for Use.

To provide context for key design decisions, throughout each section, we elaborate on participants' existing practices, challenges, desires, and needs for improving their decision-making process,

4.1 Guiding Design Principles

The goal of the guidebook is to scaffold public sector agency decision-making around the following question: *Should we move forward with developing or deploying a proposed AI tool? If yes, what are key considerations to plan for?* The guidebook aims to support agencies in answering this question through a deliberation-driven process supported by the following materials: (1) Question prompts to support conversations around the social (organizational, societal, and legal) and technical (data and modeling) considerations that should inform their recommendation, (2) Pointers to external resources to help guide their responses, (3) Template for a recommended deliverable to help communicate rationales and evidence for the recommendation that results from these deliberations, (4) Proposed use cases that illustrate how agencies could adopt the guidebook into their existing work processes, and (5) Success criteria to signal whether the intended outcomes of the guidebook may be relevant and useful to agencies.

In co-designing the guidebook towards this goal, we centered two core design principles:

- **(Design Principle 1) Promoting reflexive deliberation.** The question prompts (Section 4.2) should support stakeholders in having reflexive discussions—for example, conversations that surface their own pre-existing assumptions and beliefs about human versus AI capabilities and limitations with respect to a given task and context, or that surface relevant tacit knowledge that may be helpful to share with others. The question prompts should be designed to avoid prompting simple yes or no responses, to ensure that responses to complex questions are not reduced to a simple compliance activity. In drawing on prior work that emphasize the role of the toolkit as one that "prompts discussion and reflection that might not otherwise take place" [37, 60], this design principles extends these notions of effective toolkits from prior literature to apply to topics of importance in public sector contexts. Prior research on public sector contexts (e.g., [27, 30, 61]) as well as findings from this study suggest that agency stakeholders' differing backgrounds shape their assumptions and concerns around AI tools, motivating the need for a deliberative decision-making process that surface these individual differences. Throughout Section 4.2, we elaborate on participants' existing challenges and desires to illustrate the importance of Design Principle 1 in their contexts.
- **(Design Principle 2) Ensuring practicality of the process.** The guidebook should be designed to support a process (Section 4.3) that public sector agencies can feasibly understand, adopt, and adapt as needed. If an agency already has an existing decision-making structure, or conversations related to AI design already take place, the agency should find it easy to "fit" the guidebook into their existing organizational processes and conversations. This design principle is aimed at addressesing concerns raised in prior literature that existing responsible AI toolkits are often designed in isolation from the organizational contexts they intend to augment (e.g., [66]). This design principle is also motivated by our observations of the four public sector agencies in our study, which each had their own existing or planned organizational processes for developing AI tools (Section 4.3.1). Further, by co-designing the *process* the toolkit should follow (in addition to the guidebook content), we further an understanding of how organizational, labor, and power dynamics implicate the potential effectiveness of responsible AI toolkits in the public sector (Section 4.3.2).

4.2 Content Design: Scaffolding Reflexive Deliberation

Participants ideated critical questions that spanned four high-level topics, 12 mid-level, and 20 low-level topics. In each of the four subsections below, we briefly describe why participants were interested in the overall category of questions and provide example questions. The full set of deliberation questions for the *Situate AI* Guidebook (*Ver*.1) can be found in Appendix A.

4.2.1 *Goals and Intended Use.*

This section is intended to scaffold conversations around the following broad questions:

(1) Given our underlying goals and intended use case(s), is our proposed AI tool appropriate?
(2) What evidence do we have to support our answer to the previous question? What additional tasks may be required in the future to help us gather more evidence and/or better understand the evidence we currently have?

Sample Questions.

- Overall goal for using algorithmic tool
  - Who is going to be affected by the decision to use this hypothetical AI tool?
  - What evidence do we have suggesting that the pain-point this tool aims to solve actually exists?? What evidence do we have suggesting that technology may offer a remedy to this pain point?
  - Recall the stakeholders who are the most impacted by this hypothetical AI tool. How do we bring their voices to the table when determining goals?
  - Are there differences in the goals the agency versus community members think the tool should address? If so, what are they? If we are uncertain, what can we do to understand potential differences?
  - What biases (as a public sector agency) do we bring into this decision-making process?
- Selection of outcomes that the algorithmic tool aims to improve
  - Hypothetically, imagine that our tool does a perfect job of improving the outcome that it targets. What additional problems might this create elsewhere in the system?
- Empirical evaluations of algorithmic tool
  - Once the tool is deployed and in use, how can we evaluate how well it is working in the short-term? How can we evaluate how well it is working longer-term?
  - How can we effectively evaluate the tool from the perspective of impacted community members?
  - How might frontline workers respond to the tool? How can we better understand their underlying concerns and desires towards the tool?

The deliberation questions focus on promoting conversations that bridge reflection and understanding of the goals of the proposed AI tool, as well as how these goals will be operationalized into measurable outcomes. The 52 questions within the *Goals and Intended Use* section are divided into nine subsections: (1) Who the tool impacts and serves, (2) Intended use, (3) How agency-external stakeholders should be involved in determining goals, (4) Differences in goals between the agency and impacted community members, (5) Envisioned harms and benefits, (6) Impacts of outcome choice, (7) Measuring improvement based on outcomes, (8) Centering community needs, and (9) Worker perceptions. For the

purpose of this paper, we sample one question from each topical subsection.

Several of the questions in this section are designed to help **surface underlying assumptions regarding who benefits** from the use of the tool, and to support discussion around **what evidence suggests that these assumptions are true**. These questions stem from participants' concerns around whether their AI systems are targeting areas that would bring the most benefits, and to whom these benefits apply. For example, one participant noted that their agency had invested a lot of effort into assessing and trying to improve fairness in their algorithms. However, the participant wondered whether they should have been having conversations around larger, "more challenging" questions. For instance, they wondered whether "correcting for bias" in an algorithm within an inherently biased system is a meaningful or feasible goal. They further elaborated:

> "I think there my concern often has to do with [the] unexamined belief that an algorithm is always an improvement. […] I think [questions on broader goals and benefits are] more challenging and that people [who] are running the system may not always see […] Personally, I think there's a lot of stuff that can be done with machine learning that doesn't have to [target] decision-making at the participant level. […] But those are the kinds of questions the immediate focus [is] on. 'Oh, we're going to use this to make decisions at critical points in programs.' Those are things that to me still need to be discussed. And it may be that those conversations are happening at tables that I'm just not at." (A02)

Other participants expressed concerns for how frontline workers in their agency–the majority of who are currently not involved in early-stage conversations around the goals of the AI tool–may be **misunderstanding the intended uses and capabilities of their AI tools**. For example, one participant described that frontline workers may be concerned that the AI tools will displace them, even though their agency doesn't intend to use them to automate workers' jobs. They described:

> "There's almost like a mystique around machine learning algorithms, like there's some amazing thing that is all knowing and all seen, and therefore can predict all these different things. […] helping people [… understand] what it's able to do and not able to do, I think, is something we've struggled with" (A04).

Other questions are intended to help **forefront considerations around what additional planning and resources may be needed**, in order to adequately complete a related task in the future. For example, workers within agencies often described that involving community members in their AI design and evaluation process can be challenging, given the current lack of infrastructure to support such collaborations. However, community advocates described how involving community members is often an after-thought. One community advocate described the importance of being intentional and proactive in community engagement practices, because

> "it's easy to let that be something that gets back burning, like throughout the process to just have that be something we'll get to, and then we end up in that feedback loop where the feedback is provided but the tool is already created" (C2).

Questions in this section help promote earlier reflection and planning on how community members could be involved, so that they could conduct appropriate empirical evaluations regarding their perceptions of the AI tool.

4.2.2 ***Societal and Legal Considerations.***

This section is intended to scaffold conversations around the following broad questions:

(1) Given the societal, ethical, and legal considerations and envisioned impacts associated with the use of AI tools for our stated goals, is our AI tool appropriate?
(2) What evidence do we have to support our answer to the previous question? What additional tasks may be required in the future to help us gather more evidence and/or better understand the evidence we currently have?

Sample Questions.

- Legal considerations around the use of algorithmic tool
  - Do the people impacted by the tool have the power or ability to take legal recourse?
- Ethical and fairness considerations around the use of algorithmic tool
  - Are there differences in the goals the agency versus community members think the tool should address? If so, what are they? If you are uncertain, what are your plans for understanding potential differences?
  - Can we agree on a definition of fairness and equity in this context? What would it look like if the desired state is achieved?
  - Are fairness and equity definitions and operationalizations adequately context-specific? (For example, in the child welfare domain: children with similar profiles receive similar predictions irrespective of race?)
  - Do we understand the negative impacts of the decision made across sensitive demographic groups?
- Social and historical context surrounding the use of algorithmic tool
  - Have we recognized and tried to adjust for implicit biases and discrimination inherent in these social systems that might get embedded into the algorithm?
  - How might we clearly communicate the limitations and historical context of the data to community members?

Overall, the goal of this section is to help promote a systematic, deeper conversation on the various dimensions of social and ethical concerns relevant to the design of an AI tool. The 38 questions within the *Societal and Legal Considerations* section are divided into seven subsections: (1) Legal considerations around the use of the algorithmic tool, (2) Impacted community member needs, (3) Involving impacted communities, (4) Clarity of ethics goals and

definitions, (5) Operationalization of ethics goals, (6) Envisioning potential negative impacts, and (7) Social and historical context surrounding the use of the algorithmic tool. Again, for the purpose of the paper, we sample one question per topical subsection.

Participants shared that they did not currently have **structured opportunities to proactively discuss social and ethical considerations surrounding** AI tool design. While participants described that their teams spent a lot of time working on related data- and model-specific fairness tasks (e.g., using bias correction methods to improve the fairness of their AI tool), several participants noted a **desire to discuss normative concerns regarding the design of an AI tool that could only be addressed in earlier problem formulation stages**. Moreover, participants' past experiences illustrated an opportunity to better support cross-stakeholder communications around the ethical considerations that should aid AI design, by equipping teams with a **shared knowledge base and vocabulary** for ethical concerns. For instance, one participant described how a leadership team tasked them with creating a predictive algorithm to assist decisions about fraud investigation. The participant's team tried to "get them away from this" because the task was technically infeasible (producing high false positive rates) and ethically risky the cost of errors is high, given that decisions to investigate are highly intrusive to the individual. This section's questions intend to support agency stakeholders in forming a more complete understanding of the different ethical factors that could make a proposed AI tool design "appropriate" or "inappropriate."

We note that the guidebook does not exclusively surface societal and ethical considerations in this section; the prevalence of relevant questions included in the other three topical sections (Goals and Intended Use, Data and Modeling Constraints, Organizational Governance) reflect how social and ethical considerations are intertwined with all facets of a proposed AI tool.

4.2.3 *Data and Modeling Constraints.*

This section is intended to scaffold conversations around the following broad questions:

(1) Given the availability and condition of existing data sources, and our intended modeling approach, is our proposed AI tool appropriate?
(2) What evidence do we have to support our answer to the previous question? What additional tasks may be required in the future to help us gather more evidence and/or better understand the evidence we currently have?

Sample Questions.

- Understanding data quality
  - Has the definition of the data changed over time? (E.g., in child welfare, has reunification always meant to reunify with the parent?)
- Process of preparing data
  - How are we preprocessing the data?
  - Who should be involved in making decisions around whether to include or exclude certain data points or features? Do we have plans for involving those people?
- Model selection
  - Is our model appropriate given the available data? Why or why not?

This section intends to forefront conversations around data and technical work that may be critical to have earlier on. The 18 questions within the *Data and Modeling Constraints* section are divided into seven subsections: (1) Understanding data quality, (2) Process of preparing data, (3) Selection of statistical models to fit data. For the purpose of the paper, we provide a subsample of questions under each topical subsection.

Participants who had experience developing AI tools often underscored the importance of ensuring that they had the computing resources and data needed to develop their proposed AI tool. For example, they described the importance of forming a context-specific understanding of the data labels that may be challenging to identify without relevant domain knowledge (e.g., whether certain labels like "reunification" have changed definitions over time). Others described the importance of deliberating who should be involved in data inclusion and exclusion decisions when they are cleaning their data.

4.2.4 *Organizational Governance Factors.*

This section is intended to scaffold conversations around the following broad questions:

(1) Given our plans for ensuring longer-term technical maintenance and policy-oriented governance, do we have adequate post-deployment support for our proposed AI tool?
(2) What evidence do we have to support our answer to the previous question? What additional tasks may be required in the future to help us gather more evidence and/or better understand the evidence we currently have?

Sample Questions.

- Long-run maintenance of algorithmic tool
  - Do we expect there will be shifts in performance metrics over time? If so, why? What are our plans for identifying and mitigating those shifts?
  - Do we have the mechanisms to monitor whether the tool is having unintended consequences?
- Organizational policies and resources around the use of algorithmic tool
  - Is there training for frontline workers who will be asked to use the tool? What evidence suggests that this training is adequate?
  - Imagine that we could assemble the "ideal team" to monitor and govern the tool after it is deployed: What are the characteristics of this ideal team?
    - Who is the *actual* team that will monitor and govern the tool after it is deployed?

    - Given the gaps between the "ideal team" and the actual team we expect to have: What risks to post-deployment monitoring and governance can we anticipate? How might we mitigate these risks?
- Internal political considerations around the use of algorithmic tool
  - How well do we understand system administrators' and leadership's perspectives around the use of this tool?
  - How well do staff and leadership understand 'why' the tool could bring value?

The 24 questions within the *Organizational Governance Factors* section are divided into five subsections: (1) Measuring changes in model performance over time, (2) Mechanisms to identify long-term changes in model performance, (3) Policies around worker interactions with the AI tool, (4) Governance structures around the AI tool, and (5) Internal political considerations around the use of the AI tool. As with prior sections, we include in this paper a sample of questions across these topical subsections.

Similar to considerations around the *Social and Legal Considerations* of AI design (Section 4.2.2), participants often described encountering challenges when attempting to meet organizational governance-related needs of the AI tool, like maintaining their AI tool over time, ensuring workers are adequately trained, or communicating the goals and capabilities of the AI tool to agency leadership. Partipants highlighted that many of these challenges arise because such considerations are discussed in an ad-hoc manner, too late in the AI development process. Given that several of these needs may require longer-term planning and preparation (e.g., gathering resources of model maintenance), public sector agencies may be better equipped in meeting these governance needs if they were discussed in early stages of model design (rather than after an AI tool has already been developed). For example, participants described how they currently lack domain experts that could help maintain and improve their model post-deployment—a gap in their AI development process that they felt was critically important to address. While agencies currently discuss maintenance-related concerns at the deployment stage, this may not allow the agencies enough time to deliberate who should be involved in maintenance, or how to allocate additional roles for a maintenance team.

4.3 Process Design: Designing for Practicality and Adaptability

The overall goal of the Situate AI Guidebook is to help public sector agencies make more informed, deliberative decisions about whether and how to move forward with implementing a proposed AI tool. Prior literature studying existing responsible AI toolkits have started to surface concerns around how such toolkits may be used inappropriately or not used at all in practice, due to misalignments with the organizational contexts they are designed to support [37, 66]. In this section, we describe findings related to the broader deliberation *process* that participants envisioned the deliberation questions (Section 4.2) could be used to support.

Below, we first present our proposed use case for the *Situate AI* Guidebook, along with an example instantiation of the use case and an explanation of how participants' existing practices informed this use case. We then discuss participants' desires for alternative use cases and processes around deliberation. Participants across agencies and roles expressed interest in using the questions in a few different ways, based on their concerns around cross-stakeholder power dynamics and desires to enable deliberation practices that align with their organizational values [51, 66]. Given participants' interests in adapting the guidebook to different use cases, a key component of the *Situate AI* Guidebook is that it is designed to allow users to select which topics and questions they would like to focus on: The deliberation questions are categorized and grouped into modular components; and users have the flexibility to select from a large set of deliberation questions within each component to identify a subset that is most relevant to their use case.

4.3.1 *Proposed Use Case: Using the Guidebook to Support Structured and Iterative Deliberations.*

Participants envisioned that the guidebook could be effectively used to support structured, iterative deliberation through formal workshops between members of their agency. In this section, we elaborate on one possible way this use case can be instantiated into an overall deliberation process, then discuss how this compares to participants' existing practices within their agency. We provide an example of one possible implementation of a formal deliberation process, using the guidebook.

Example instantiation of proposed deliberation process. This process involves a four-stage phase, where the public sector agency would first appoint a facilitator(s) to help organize the overall decision-making process:

Stage 1: Topic and Attendee Identification. The facilitator will identify which of the four guidebook topics, if not all, they are interested in convening a deliberation workshop on. Based on broad guidance provided in the guidebook, the facilitator will then identify the stakeholders that should be included in the deliberation workshop based on the selected topics. For example, the Societal and Legal Considerations section is designed to be used by a more diverse range of stakeholders (e.g., AI practitioners, frontline workers, community members, legal experts) compared to the Data and Modeling Constraints section (e.g., AI practitioners only).

Stage 2: Question Selection and Deliberation. After finding a shared time for the deliberation workshop, the facilitator should share the goals and topics of deliberation (included in the guidebook) with the group. The guidebook includes a large number of questions for each topic of deliberation. For example, the *Goals and Intended Use* section alone has 52 questions. To ensure that the questions can be feasibly discussed within the allocated time, we highlight 1-2 recommended questions per major subsection, resulting in a smaller number of questions (19 questions for the *Goals and Intended Use* section). The remaining questions are also available in the guidebook as "optional questions." We recommend that, at the start of the workshop, the facilitator provides the attendees with the opportunity to identify any questions from the "optional questions" category they would like to additionally or alternatively discuss in the workshop. As the attendees are discussing each question, the facilitator should take note of their responses and points of

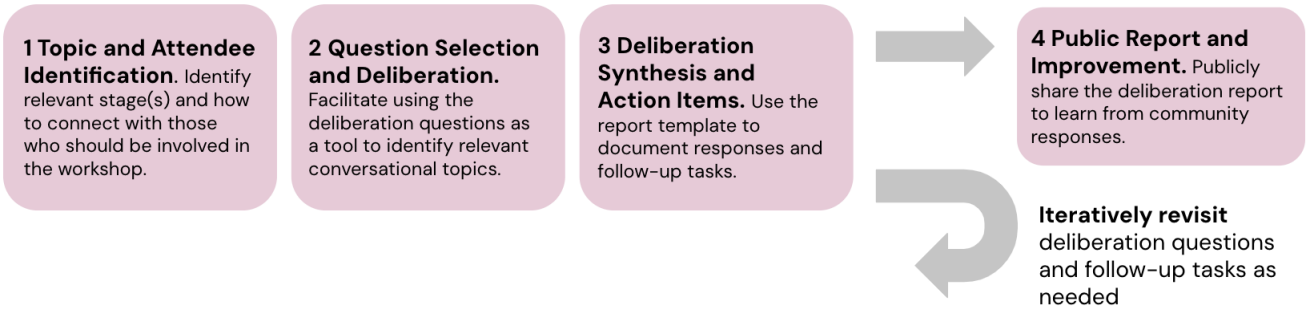

Figure 4: A high-level overview of the main stages involved in the proposed use case for the *Situate AI* Guidebook, intended to support structured and iterative deliberations within a public sector agency.

disagreement. If there are disagreements that are challenging to resolve in response to a question, the facilitator should help the group identify action items to help gather more information or perspectives and plan to revisit the question at a later time. If the group finds they currently lack the resources or knowledge to fully address a question, the facilitator should also make note of this and plan to revisit the question at a later time.

Stage 3: Deliberation Synthesis and Action Items. After the deliberation workshop, the facilitator should summarize the discussions and outcomes into the deliberation report template which we include in the guidebook. The template includes the following questions: *(1) What is your recommendation? (2) Please list core reasons for your recommendation, based on the deliberation workshop, (3) Are there any follow-up tasks you must complete, in order to fully support this recommendation? If yes, please write the task(s) and plan(s) for completing it, and (4) What core counter-arguments against this recommendation arose during the deliberation workshop? Please describe each counter-argument, including how you addressed or plan to address each one.* Based on the report responses, the facilitator should continue to iteratively revisit the deliberation questions, organizing additional deliberation workshops with the attendees as needed, as they complete the follow-up tasks included in the report.

Stage 4: Public Report and Improvement. The facilitator should work with their agency to publicly share the deliberation report with agency-external stakeholders, including impacted community members and related organizations. To promote conversations and bidirectional learning between community members and agency-internal stakeholders, the agency should hold public convenings and host an online commenting forum for any individuals who would feel more comfortable contributing anonymously online. The agency should then synthesize the themes that emerged from the conversations, identify action items to address any concerns, and share these results of the community conversations with the public. In the guidebook, we plan to include links to existing community review efforts to provide examples of what this interaction could look like. However, we note that effectively completing this step requires additional research and resource creation (which we discuss in the Discussion Section 5).

How participants' existing organizational practices informed the proposed process design. Reflected in the process above, participants raised several important considerations to ensure the process is practical and meaningful to their agency. For example, participants in agencies that are actively developing new AI tools described that there are already AI design and development processes in place that support focused discussions on improving, for example, the algorithmic fairness of their AI tool. These participants did not want the Situate AI Guidebook to replace these conversations and work sessions. Instead, they desired a process that could **augment their existing processes**. For instance, participants noted that having these deliberation workshops earlier on, before they developed or analyzed any AI model, can help promote reflexive conversations about what it means to do the work of AI fairness and what important considerations that should aid this work (e.g., whether there is a definition of "fairness" that agency workers agree on). Relatedly, as reflected in the process description, several participants described the importance of **revisiting these deliberation questions iteratively**, throughout the AI development and deployment process (rather than only discussing these at the early ideation stages of a development process). For instance, participants described that their understanding of some of the question responses (e.g., the intended outcomes that the AI tool should help achieve) may evolve with time as the AI tool development evolves (e.g., depending on what is technically possible, given data constraints or prediction errors). Participants noted that designing the guidebook to center an iterative process would help ensure the conversations complement their existing AI development process, which is also iterative in nature. We originally designed the process so that the deliberation workshop attendees would be required to come to a consensus on the final recommendation before moving forward. However, the participants we spoke with emphasized that this may be impractical and unnecessary based on their end goals for the guidebook. We elaborate on this point in Section 4.4 when we discuss the guidebook's Success Criteria.

We discuss limitations and future work related to this proposed use case in the Discussion, drawing on prior literature suggesting ways in which research-based conversational tools may not be adopted in practice or may be used in inappropriate ways.

4.3.2 ***Empowering Participation: Accounting for Organizational Power Dynamics.*** Findings from this study strongly suggest that frontline workers—those who would be asked to use the AI tool once deployed—are interested in engaging in early-stage deliberations around what the AI tool should be designed to assist. Prior literature also suggests that ensuring agency leadership and AI developers understand frontline workers' needs and challenges is critical to ensuring that the "right" AI tools are being developed (e.g., [29, 30, 58, 67]). However, effectively supporting conversations between roles with prominent and knowledge differentials remains a challenging task [27]. For the current version of the *Situate AI* Guidebook, we begin accounting for this challenge by editing the language used in some of the guidebook sections to ensure it is understandable to those without prior knowledge on technology. We additionally asked frontline workers about their desires for how they would want to participate in the deliberation process, to ensure they feel safe to share any concerns. In this section, we discuss these findings. However, we note that future work is needed to ensure that the *Situate AI* Guidebook ($Ver.1.0$) adequately accounts for organizational power dynamics. In the Discussion, we discuss implications for complementary policy interventions that may be needed for an agency to effectively facilitate a multi-stakeholder deliberation process.

Frontline workers' preferred processes for participating in multi-stakeholder deliberations. Frontline workers in our study had a range of perspectives around how best to involve them and their colleagues in the deliberation process. One frontline worker suggested that their agency should require all frontline workers to attend the deliberation workshops. They described that, without making participation in these discussions mandatory, frontline workers may opt to skip meetings given their busy schedules. The participant further expressed concerns that, if participation was on a voluntary basis, frontline workers who join may be harmfully self-selective: "... particularly for people with marginalized identities, it's important for them to be a part of these spaces and voice their concerns. I think that if it wasn't mandatory, it might be, you know, [a] self-selecting group" (F7). This frontline worker, along with another worker, also described how adjustments to the proposed process could help frontline workers feel more comfortable raising concerns. For example, the participants expressed that, for multi-stakeholder conversations, some frontline workers may feel more comfortable having a separate frontline worker-only deliberation workshop, synthesizing and formalizing their perspectives, then going to a group meeting with other agency stakeholders to present their perspectives. As the participant described:

> "A lot of social workers are very non-confrontational [...] we are our clients' best advocates but not for ourselves. And so I definitely do think that people might be more comfortable, you know, having their own sort of peer group discussion or colleague group discussion. And then that being you know, sort of the concerns being written down and formalized, and that being taken rather than a more informal sort of like anyone who has concerns just raise their hand and say their piece. I feel like that might be a bit daunting for some people" (F7).

Alternative use case: Using the guidebook to empower everyday conversations. Complementing frontline workers' desires for engagement, participants from agencies that were not yet developing new AI tools described a different use case for the Situate AI guidebook. These participants envisioned that the guidebook could be used by teams to support everyday conversations, with the aim of proactively avoiding pitfalls in AI project ideation and selection. For instance, one participant described that they wanted the guidebook to be used more casually, by everyone in the agency, to help all staff members feel empowered to "be able to do a little more of the innovation" (L7). The participant described that, even if they get stuck or need help, "it would be awesome for them to have a library of resources that they can look at" to help them get started. The participant further described that workers in their agency should "have the flexibility to structure the deliberation workbook to their needs," for example, deciding which questions to discuss, how much time to take in discussing the questions, and who to talk with. By having a guidebook that empowers any staff member to discuss topics around AI, the participant hoped that these deliberations could have rippling effects on their agency's overall culture:

> "Maybe we're trying to get from [...] 'let's do this big project right with the right leaders and things in the room' to 'how do we create a culture of improvement' [...] Not just how do we do a technology project the right way, but actually can it have a broader impact on culture. This is how we do anything. It's always with this batch of questions in mind and thinking about how we can be people around problem solving" (L7).

Desires to expand participation to community members for future versions of the guidebook. Several participants across agencies and community advocacy groups were interested in involving community members in the deliberation workshops supported by the guidebook. Workers within the agency wanted guidance on how to do this effectively. In our conversations with community advocates, we probed on how they would like to be involved in deliberations around the Situate AI guidebook. They described the importance of compensating community members for their time, providing multiple channels for communication (e.g., online forums and in-person meetings), and following up with the outcomes of the conversations: "that happens a lot, you know–agencies are like 'oh, we engage with the community, and we brought them into the space with us.' But then there's no follow up or follow through from those conversations. And that's been a historical thing" (C2).

Importantly, we note that the guidebook, in its current form, is designed to support conversations across workers within a public sector agency. It is not designed to directly support conversations between agency-internal workers and agency-external stakeholders (e.g., community members). In the Discussion section, we discuss opportunities to expand participation through design improvements.

4.4 Success Criteria

What outcomes do public sector agencies and impacted community members consider "meaningful," when assessing the effectiveness of the *Situate AI* Guidebook? What are their underlying theories of change around how their public sector agency could move towards

more responsible early-stage AI design practices, and how do they envision the *Situate AI* Guidebook can help them progress towards that path? Overall, through the guidebook, participants wanted to form an understanding of the disagreements and tensions across agency workers felt most strongly about, to help position themselves to better address these disagreements through changes to the problem formulation or design of a proposed AI tool. Importantly, as indicated by the process design in Section 4.3, participants described that the goal of the Situate AI Guidebook should *not* be to *resolve* these tensions and disagreements across individuals. Participants described that this is an infeasible task, given underlying differences in values and goals across agency stakeholders. In this section, we elaborate on four success criteria of the guidebook. These success criteria intend to help communicate the intended goals and boundaries of the guidebook, and including how they are informed by participants' own notions of success for deliberations around the design of AI tools.

4.4.1 ***Make it easier for different agency stakeholders to communicate with each other about AI design, evaluation, and governance considerations.*** Many of the challenges that participants in our study described could be addressed if better communication channels existed between different agency stakeholders—including amongst AI developers, agency leadership, and frontline workers. This challenge is also well-documented in prior literature studying public sector AI decision-making [27, 29, 31, 58]. For instance, in current practice, frontline workers are often not meaningfully involved in early-stage deliberations around AI design and adoption. As a result, agency leadership and AI developers have interpreted workers' concerns around AI as a signal for not understanding what the AI tool does. Involving frontline workers in these earlier discussions can both help more proactively inform workers of AI capabilities and empower them to engage in constructive conversations that would improve the design of AI tools.

4.4.2 ***Bring context-specific needs for resources to the forefront of AI project selection conversations.*** While participants often knew which resources they needed to successfully implement a given AI tool, their past challenges sometimes surfaced a missed opportunity to identify these needs at an earlier stage of their AI design process. Moreover, related to the previous success criterion, our conversations with the participants surfaced ways in which having agency workers with different roles and perspectives engaged in these early-stage deliberations can strengthen their ability to anticipate the potential impacts of AI design decisions. For example, frontline workers voiced that AI tools they had used in the past were designed in ways that conflicted with their existing decision-making policies; other workers described that AI deployments may add additional labor to their day-to-day tasks, given they may be asked to more diligently collect data. In current AI development processes, where frontline workers may only be meaningfully engaged in AI implementation or piloting stages, mitigating these negative impacts may require more substantive tasks like redesigning the AI tool. Participants further described that these resource-related needs were highly context-specific. For instance, when discussing the importance of anticipating how their AI tool may impact community members, one participant recalled how even the definition of "community" may differ across agencies and AI tools: "Because we would always say, 'we're doing stuff [where] the community is informing us.' And then we realized, 'oh wait, it wasn't necessarily the group of people who were impacted by [our decisions]'" (L7).

4.4.3 ***Make social and ethical considerations a first order priority in conversations around whether to move forward with an AI tool idea.*** As described in Section 4.2.2 and 4.2.4, participants described their past assessments around whether an AI tool was appropriate to implement centered algorithmic considerations—whether that be the quality of their training data or outcomes of algorithmic fairness or accuracy metrics. While these considerations are critically important, others also discussed a desire to rigorously deliberate the underlying values embedded in design decisions, and the social and ethical impacts of a proposed AI tool. This echoes concerns from prior literature, discussing how technical design decisions often include hidden policy decisions and value judgements [20, 61]. Prior work also suggests the importance of these considerations, noting that existing AI ethics toolkits have largely framed the work of ethics as "technical work" [66].

4.4.4 ***Make "fitting" an AI tool into a workplace a design problem, rather than an implementation problem.*** This success criterion intends to avoid practices where the AI tool idea is conceived before fully understanding context-specific practices and needs, and in turn, creating AI tools that frontline workers must then attempt to "fit" into their existing workflow. This tendency to treat "fitting" AI tools as an implementation problem, and its negative impacts on workers' ability to improve their existing decision-making practices, is also well documented in prior literature [67, 68]. Indeed, participants—including both AI developers and frontline workers-described that they wished they could have had better conversations, early on, to understand what the actual goal of the tool they were building should be. As one AI developer described, recalling a past experience in their team where leadership had asked them to create an AI tool: "It was kind of hard to get a sense of what the actual issue was that was being asked to be solved. It sounds kind of a lot like, 'Here's a bunch of different potential places an algorithm might fit in'" (A04). Ultimately, they were asked to create a predictive model to "to find fraud where there wasn't already suspicion of fraud." However, the AI developer described feeling leadership had proposed the idea as "this cool thing we could do" but, in reality, realizing that creating such a tool would create more problems downstream in the system (in this case, it would create too many referrals to be able to investigate). By promoting early-stage, structured deliberations around critical topics related to AI tools, public sector agencies could be supported in identifying higher-value, lower-risk opportunities to innovate with AI systems.

5 DISCUSSION

Public sector agencies in the U.S. are increasingly exploring how new AI tools can assist or automate services in child welfare, homelessness housing, healthcare, and policing, among other domains [11, 22, 45, 58, 67]. In the U.S., these public services have historically been characterized by racial inequity, procedural injustice, and distrust from the impacted communities [9, 17, 61]. While agencies

have rapidly begun to deploy AI tools to improve their services, ensuring responsible development has proven to be an immense challenge. In the past decade, such AI tools have often failed to serve the needs of the communities that agencies are expected to serve [6, 17, 20, 24, 25]. A growing body of literature has recognized that many downstream harms resulting from AI tools can be traced back to decisions made during the earliest problem formulation and ideation stages of the AI lifecycle. Yet, there are few, if any, effective resources for public sector agencies in making more deliberate decisions regarding *whether* a given AI proposal should be developed in the first place.

Through iterative co-design sessions with 32 individuals (agency leaders, AI developers, frontline decision-makers, and community advocates) across four public sector agencies and three community advocacy groups, we created the *Situate AI* Guidebook ($Ver.1.0$). The guidebook, designed for public sector agency workers, scaffolds the process for early-stage deliberations around *whether and under what conditions* to move forward with the implementation or adoption of a new AI tool or idea. To support this process, the guidebook presents a set of 132 deliberation questions–which participants indicated are critical to consider yet are often overlooked today–spanning both social (organizational, societal, and legal) and technical (data and modeling) considerations around AI; along with guidance on the overall deliberative decision-making steps and success criteria for use. In this section, we discuss the design decisions we made in creating the guidebook, along with limitations and opportunities for future work. For each section of this discussion, we begin by summarizing relevant portions of the findings. Then, we elaborate on limitations and future opportunities to improve upon the existing guidebook.

5.1 Overcoming Low Adoption Rates for Responsible AI Toolkits in the Public Sector

As the research community continues to innovate new Responsible AI toolkits, recent literature has raised concerns regarding the practical efficacy of such toolkits. Prior work has found that the majority of AI ethics toolkits fail to account for the relevant organizational context, hindering their usability (e.g., overlooking guidance on *how* different stakeholders should be engaged) and effectiveness (e.g., focusing on the technical but neglecting the social aspects of AI ethics work) [66]. Public sector decision-making around service allocation is often shaped by resource and staffing shortages, and require balancing tradeoffs to meet the competing needs of a range of stakeholders (e.g., impacted community members, policymakers and regulators, politicians) [64]. Moreover, AI tools in the public sector often target socially high-stakes decisions (e.g., whether to screen in a family for child maltreatment investigation, or provide an individual with a credit loan), that have disproportionately negatively impacted the lives of historically marginalized communities. Prior work has shed light on the downstream impacts that public sector AI systems have had (e.g., [9, 55]), along with challenges to ensuring their responsible design and use (e.g., [30, 58, 64]). Through our study, we demonstrated how collaborating with public sector agencies and community members to co-design a responsible AI toolkit–including its process and content design–can help surface and account for some of these challenges. That said, future research is needed to understand how effective the toolkit is in practice, and to surface other challenges that can only be observed through actual use (rather than through our co-design and interview study format). In the following subsections, we briefly discuss related findings and opportunities to improve the contextual design and use of the Situate AI Guidebook for public sector settings.

5.1.1 Designing for more inclusive forms of worker participation. While AI tools for public sector contexts implicate a range of different agency-internal stakeholders, these agency workers—from agency leaders to frontline workers—often operate in silos, separated by power imbalances and knowledge differentials. We found that participants desired a range of particiation structures to account for these differences. For example, some frontline workers wanted to first gather amongst others with similar occupations to prepare for the deliberation workshop, and then send in one frontline worker to attend the workshop and represent their perspectives. On the other hand, other participants suggested that there should be an organizational policy that required all frontline workers to attend the deliberation workshops alongside the AI developers and agency leaders. Future work is needed to understand the broader range of solutions that could best address these differences in workers' preferences. For example, future work could pilot different processes, where agency workers are grouped in deliberation workshops in specified configurations depending on their role and background. Through observations and retrospective interviews of these configurations, we could better understand whether having a set of deliberation questions alone is adequate to prompt meaningful conversations. Future work could additionally explore how additional resources and toolscould be used alongside the deliberation toolkit, in order to effectively scaffold conversations around the appropriateness of AI design ideas. This direction would be especially critical to pursue, in order to ensure that the deliberation toolkit is accessible to those who may not have had prior exposure to AI technologies.

5.1.2 Incentivizing and governing responsible use. Ensuring responsible use and adoption of the *Situate AI* Guidebook may require complementary efforts from governing bodies. For example, while the U.S. does not currently require public sector agencies to document early-stage deliberations around AI, having similar forms of external forces that incentivize agencies to engage in early-stage deliberation may help ensure that the deliberation toolkit is used effectively. One way to incentive public sector agencies may be to clearly communicate how the toolkit aligns with and complements existing voluntary guidelines, such as those in the NIST AI Risk Management Framework (RMF) [44]. While the NIST RMF and NIST RMF Playbook [43] both focus on providing higher level guidance on steps to follow for responsible AI design, research-based co-designed toolkits like the Situate AI Guidebook can help bridge gaps between their proposed policy guidance and real-world practice. In future work, we plan to map the guidebook to the four functions captured in the AI RMF Core: Govern, Map, Measure, and Manage. For example, each question or category of questions could be assigned one or more of the AI RMF Core functions.

In future work, we plan to explore with public sector agencies community advocates, and other stakeholders how new policy and organizational interventions can support them in using the *Situate*

AI Guidebook. The public sector agencies in our study, including the frontline workers, expressed interest in exploring how to use the guidebook in practice through pilots.

5.2 Expanding the *Situate AI* Guidebook to Engage Community Members

In public sector contexts, there is often a greater expectation that decisions center the needs of the community, including by being transparent to and engaging with the community during the decision-making process. In our study, participants expressed a desire for guidance on how to engage with community members in discussing complex topics around AI design. While the deliberation protocol is not currently designed to support such conversations, the current version poses questions that suggest follow-up tasks involving conversations with community members. For example, the question "Are we assessing the tool from the perspective of impacted community members? What evidence do we have to suggest that we are genuinely understanding their concerns and desires?" suggests that the agency should talk with impacted community members to understand their perspectives–a task that would require additional guidance and resources to complete successfully. Agency workers acknowledge they are often pushed to involve community members in their AI design work but without actionable guidance on how to do so effectively. Participants suggested linking existing relevant resources from the guidebook, to assist agencies in this regard. Moreover, community advocates in our study expressed interest in engaging in the deliberation workshops themselves.

Future work should explore ways to improve the design, structure, and process of the deliberation guidebook so that it is well-equipped to support conversations between agency-internal workers and agency-external community representatives. For example, to help bridge a shared vocabulary about AI between agency workers and community representatives, future work could begin by integrating existing resources and guidance from publicly available guides like *A People's Guide to Tech* [48]. It is possible that the specific questions and topics this deliberation guidebook addresses requires additional scaffolding and support. Future work should explore ways to provide this support through continued collaborations with community advocacy groups. Community advocates in our study additionally expressed interest in having both the option to attend in-person workshops and to participate anonymously online. Future work could explore ways to support more democratic forms of participation online using social computing platforms (e.g., [34]) intended to facilitate and analyze deliberation about specific topics around AI.

5.3 Exploring how the *Situate AI* Guidebook Can Support Deliberation in Non-Public Sector Contexts

While the *Situate AI* Guidebook was originally designed for high-stakes public sector decision-making domains, there is an opportunity to adapt it to meet the needs of other AI use cases. Private and public sector settings share many organizational challenges (e.g., communication barriers across teams and occupations) and development tendencies (e.g., targeting problem spaces that AI capabilities may be ill-suited towards), that could implicate the effective design of responsible AI toolkits. Moreover, by designing for a setting with relatively high expectations and standards for responsible design (i.e., *high stakes* AI applications in the *public sector*), the *Situate AI* Guidebook sets a high bar for the kinds of questions and processes that should be followed to responsibly evaluate early-stage AI design concepts elsewhere. For this reasons, we expect that the guidebook may be (at least partially) applicable to other AI deployment contexts, including certain high risk applications in industry (e.g., healthcare, credit lending).

Indeed, many of the questions that the participants generated are relevant to AI deployment contexts beyond the public sector. Most deliberation questions target core issues relevant to all AI deployments (i.e., around the goals, ethical implications, technical constraints, and governance practices surrounding an AI deployment). Besides the deliberation questions, the design principles underlying the guidebook may also help make the guidebook useful for non-public sector contexts. Because the public sector agencies we partnered with differed in their organizational practices (e.g., who is involved in decision-making around AI) and priorities (e.g., types of services provided), we intentionally designed the guidebook to allow for flexible adoption and personalization. For instance, in designing towards Design Principle 2, we categorized the questions into modular topics and subtopics that can be selected and combined for a given deliberation workshop. There is an opportunity for future work to expand the set of labels attached to the deliberation questions. For example, participants described an interest in future versions of the toolkit that categorized questions based on the type of technology (e.g., generative AI, predictive analytics), or the type of deliberation (e.g., individual assumption-checking, knowledge-sharing, future task identification) that the question is intended to support. Future work could similarly aim to understand the types of questions that are the most critical for certain AI deployment contexts (e.g., public sector social work, private sector healthcare, etc.).

6 CONCLUSION

As public sector agencies in the U.S. increasingly turn to AI tools to increase the efficiency of their services, it becomes critical to ensure these tools are designed responsibly. While much research and development efforts have been dedicated to better scaffolding responsible AI development and evaluation practices, real-world AI failures often point to fundamental problems in the problem formulation of an AI tool–problems that should be addressed before proceeding with any decision to develop an AI tool. Yet, we currently lack effective processes to support such early-stage, deliberate decision-making in the public sector. This paper introduces the *Situate AI* Guidebook: the first toolkit that is *co-designed* with public sector agencies and community advocacy groups to scaffold *early-stage deliberations* regarding *whether or not* to move forward with the development of an AI design concept. Through co-design sessions conducted over the course of 8 months, participants generated 132 questions which we organized under four high-level categories including (1) *goals and intended use*, (2) *social and legal considerations* of a proposed AI tool, (3) *data and modeling constraints*, and (4) *organizational governance factors*. In this paper, we elaborate on how participants' practices, challenges,

and concerns shaped the *Situate AI* Guidebook's guiding design principles, the deliberation questions they believed were critical for early-stage decision-making, the overall organizational and team decision-making process the guidebook should scaffold, and the success criteria used to assess the effectiveness of the guidebook. We additionally discuss opportunities for future work to improve the design and implementation of the *Situate AI* Guidebook, including via continued partnership with public sector agencies in our study, who plan to pilot how the guidebook can be used in their agency.

ACKNOWLEDGMENTS

We thank the public sector agencies and community advocacy groups that shared their time, perspectives, and expertise with us to help create this paper and guidebook. We also thank the anonymous reviewers for their thoughtful feedback that helped improve this paper. The researchers gratefully acknowledge the support of the Digital Transformation and Innovation Center at Carnegie Mellon University sponsored by PwC. This research was also supported by funding from the UL Research Institutes (through the Center for Advancing Safety of Machine Intelligence), the National Science Foundation (NSF) (Award No. 1952085, IIS2040929, and IIS2229881), the NSF Graduate Research Fellowship Program, and the K&L Gates Presidential Fellowship (through Carnegie Mellon University). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of NSF or other funding agencies.

REFERENCES

[1] 2022. Community jury - Azure Application Architecture Guide. https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/

[2] 2023. Algorithmic Impact Assessment tool | Government of Canada. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

[3] 2023. A guiding framework to vetting public sector technology vendors. https://www.fordfoundation.org/work/learning/research-reports/a-guiding-framework-to-vetting-public-sector-technology-vendors/

[4] J Khadijah Abdurahman. 2021. Calculating the Souls of Black Folk: Predictive Analytics in the New York City Administration for Children's Services. In *Colum. J. Race & L. Forum*, Vol. 11. HeinOnline, 75.

[5] Accenture. 2018. Public services in the era of artificial intelligence. https://www.accenture.com/us-en/services/public-service/artificial-intelligence

[6] Yang Bao, Gilles Hilary, and Bin Ke. 2022. Artificial intelligence and fraud detection. *Innovative Technology at the Interface of Finance and Operations: Volume I* (2022), 223–247.

[7] John Billings, Ian Blunt, Adam Steventon, Theo Georghiou, Geraint Lewis, and Martin Bardsley. 2012. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). *BMJ open* 2, 4 (2012), e001667.

[8] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[9] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[10] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*. 1–22.

[11] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. PMLR, 134–148.

[12] Alexandra Chouldechova, Emily Putnam-Hornstein, Suzanne Dworak-Peck, Diana Benavides-Prado, Oleksandr Fialko, Rhema Vaithianathan, Sorelle A Friedler, and Christo Wilson. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research* 81 (2018), 1–15. http://proceedings.mlr.press/v81/chouldechova18a.html

[13] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2022. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. *arXiv preprint arXiv:2206.14983* (2022).

[14] Tim Dare and Eileen Gambrill. 2017. Ethical analysis: Predictive risk models at call screening for Allegheny County. *Allegheny County Analytics* (2017).

[15] Glyn Elwyn, Isabelle Scholl, Caroline Tietbohl, Mala Mann, Adrian GK Edwards, Catharine Clay, France Légaré, Trudy van der Weijden, Carmen L Lewis, Richard M Wexler, et al. 2013. "Many miles to go…": a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC medical informatics and decision making* 13 (2013), 1–10.

[16] David Freeman Engstrom, Daniel E Ho, Catherine M Sharkey, and Mariano-Florentino Cuéllar. 2020. Government by algorithm: Artificial intelligence in federal administrative agencies. *NYU School of Law, Public Law Research Paper* 20-54 (2020).

[17] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

[18] Aline Shakti Franzke, Iris Muis, and Mirko Tobias Schäfer. 2021. Data Ethics Decision Aid (DEDA): a dialogical framework for ethical inquiry of AI and data projects in the Netherlands. *Ethics and Information Technology* (2021), 1–17.

[19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[20] Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov, Kath Xu, Anjana Samant, and Aaron Horowitz. 2023. The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1292–1310.

[21] Karine Gravel, France Légaré, and Ian D Graham. 2006. Barriers and facilitators to implementing shared decision-making in clinical practice: a systematic review of health professionals' perceptions. *Implementation science* 1 (2006), 1–12.

[22] Ben Green. 2020. The false promise of risk assessments: epistemic reform and the limits of fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 594–606.

[23] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45 (2022), 105681.

[24] Sally Ho and Garance Burke. 2022. An algorithm that screens for child neglect raises concerns. *AP News* (2022).

[25] Sally Ho and Garance Burke. 2023. Child welfare algorithm faces Justice Department scrutiny. *AP News* (2023).

[26] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2017. Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. In *Proceedings of the seventh international learning analytics & knowledge conference*. 257–266.

[27] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. 2020. Shifting concepts of value: Designing algorithmic decision-support systems for public services. *NordiCHI* (2020), 1–12. https://doi.org/10.1145/3419249.3420149

[28] NPR June. 2022. Oregon is dropping an artificial intelligence tool used in child welfare system. *Link: https://www. npr. org/2022/06/02/1102661376/oregon-drops-artificial-intelligence-child-abuse-cases* (2022).

[29] Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. 2023. Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use. (2023).

[30] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[31] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference*. 454–470.

[32] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[33] PM Krafft, Meg Young, Michael Katell, Jennifer E Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, et al. 2021. An action-oriented AI policy toolkit for technology audits by community advocates and activists. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 772–781.

[34] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of*

the ACM 2012 conference on Computer Supported Cooperative Work. 265–274.
[35] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
[36] Karen Levy, Kyla E Chasalow, and Sarah Riley. 2021. Algorithms and Decision-Making in the Public Sector. *Annual Review of Law and Social Science* 17 (2021), 1–38.
[37] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[38] Shefeh Prisilia Mbuy and Marco Ortolani. 2022. Algorithmic Impact Assessment for an Ethical Use of AI in SMEs. https://doi.org/10.14236/ewic/HCI2022.34
[39] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
[40] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. Assembling accountability: algorithmic impact assessment for the public interest. *Available at SSRN 3877437* (2021).
[41] J Murray. 2021. Public sector vs. private sector: What's the difference. *The Balance Small Business* (2021).
[42] Arvind Narayanan. 2019. How to recognize AI snake oil. *Arthur Miller Lecture on Science and Ethics* (2019).
[43] NIST. 2023. NIST AI RMF Playbook. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
[44] NIST. 2023. NIST Risk Management Framework. https://csrc.nist.gov/projects/risk-management/about-rmf
[45] Ziad Obermeyer and Ezekiel J Emanuel. 2016. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine* 375, 13 (2016), 1216.
[46] Laura E Panattoni, Rhema Vaithianathan, Toni Ashton, and Geraint H Lewis. 2011. Predictive risk modelling in health: options for New Zealand and Australia. *Australian Health Review* 35, 1 (2011), 45–51.
[47] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*. 39–48.
[48] Allied Media Projects. 2018. A People's Guide To Tech. https://alliedmedia.org/resources/peoples-guide-to-ai
[49] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1776–1826.
[50] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.
[51] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.
[52] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithmic Impact Assessments: A Practical Framework for Public Agency. *AI Now* (2018).
[53] John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. 2020. A methodology for creating AI FactSheets. *arXiv preprint arXiv:2006.13796* (2020).
[54] Dorothy Roberts. 2022. *Torn apart: how the child welfare system destroys black families–and how abolition can build a safer world.* Basic Books.
[55] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling assumptions clash with the real world: Transparency, equity, and community challenges for student assignment algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[56] Anjana Samant, Aaron Horowitz, Kath Xu, and Sophie Beiers. 2021. Family surveillance by algorithm: The rapidly spreading tools few have heard of. *American Civil Liberties Union (ACLU)* (2021). https://www.aclu.org/sites/default/files/field_document/2021.09.28a_family_surveillance_by_algorithm.pdf
[57] Anjana Samant, Aaron Horowitz, Kath Xu, and Sophie Beiers. 2021. Family surveillance by algorithm: The rapidly spreading tools few have heard of. American Civil Liberties Union (ACLU)(2021).
[58] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the US child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[59] Devansh Saxena, Erina Seh-Young Moon, Aryan Chaurasia, Yixin Guan, and Shion Guha. 2023. Rethinking" Risk" in Algorithmic Systems Through A Computational Narrative Analysis of Casenotes in Child-Welfare. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
[60] Katie Shilton. 2013. Values levers: Building ethics into design. *Science, Technology, & Human Values* 38, 3 (2013), 374–397.
[61] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1162–1177.
[62] Marie Utterberg Modén, Martin Tallvid, Johan Lundin, and Berner Lindström. 2021. Intelligent tutoring systems: Why teachers abandoned a technology aimed at automating teaching processes. (2021).
[63] Rhema Vaithianathan, Tim Maloney, Emily Putnam-Hornstein, and Nan Jiang. 2013. Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American journal of preventive medicine* 45, 3 (2013), 354–359.
[64] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.
[65] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2022. Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy. *Available at SSRN* (2022).
[66] Richmond Y Wong, Michael A Madaio, and Nick Merrill. 2023. Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–27.
[67] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
[68] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, et al. 2023. Creating Design Resources to Scaffold the Ideation of AI Concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2326–2346.
[69] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People+ AI Guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

A DELIBERATION QUESTIONS

This section includes the full list of questions included in the *Situate AI* Guidebook[2] (*Ver*.1.0).

A.1 Goals and Intended Use

The set of questions below are intended to support conversations around the following broader question: **Given our underlying goals and intended use case(s), is our proposed AI tool appropriate?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: Agency leadership, AI practitioners, frontline workers, community members.

A.1.1 ***Overall goal for using algorithmic tool.***

Who the tool impacts and serves.

- Who is going to be affected by the decision to use this hypothetical AI tool?
  - Who is going to be the most impacted?
- Who benefits from the use of the tool?
  - To what extent are the targeted outcomes intended to benefit the agency, versus the community?

Intended use.

- What evidence do we have suggesting that the painpoint this tool aims to solve actually exists?
- What evidence do we have suggesting that technology may offer a remedy to this painpoint? (Evidence may include, for

[2]https://annakawakami.github.io/situateAI-guidebook/

example, historical agency metrics, legislature, community members, research reports.)
  - What evidence suggests the specific form of technology we are envisioning (e.g., predictive analytics) may offer a remedy?
- What are the additional challenges and risks associated with pursuing a technological solution to this problem?

Involving agency-external stakeholders in determining the goals.

- Think about the most impacted stakeholders you identified in response to the questions above. How do we bring their voices to the table when determining goals?
- How can we open opportunities for those who are most impacted by the new tool to inform the decision-making process?
- When will we start to engage impacted communities in discussions around how the tool should be designed or used?

Differences in goals.

- Are there differences in the goals the agency versus community members think the tool should address? If so, what are they? If we are uncertain, what can we do to understand potential differences?
- What evidence do we have that we adequately understand the outcomes the community cares about?
- To what extent are we optimizing the things the agency cares about versus what impacted community members care about?
- Is the process we have in mind for achieving a community-oriented outcome (e.g., child safety) also aligned with the community's desires?

Envisioned harms and benefits.

- What are the potential harms and benefits of the tool, and to whom?
  - Do benefits outweigh the harms?
  - Do we expect there to be tradeoffs between accuracy, fairness, explainability? For example: making decisions in a completely random fashion may look "fair", but is not necessarily accurate.
  - Will this tool help us better allocate (scarce) resources?
- What biases (as a government agency) do we bring into this decision-making process?
  - How can we identify and mitigate them? What forms of collaboration (e.g., with impacted community members) can help us do this?
- How does this tool help us better deliver to the people we are serving, if at all?

A.1.2 ***Selection of outcomes that the algorithmic tool aims to improve.***

Impacts of outcome choice.

- Hypothetically, imagine that our tool does a perfect job of improving the outcome that it targets. What additional problems might this create elsewhere in the system?
- To what extent are we optimizing the things the agency cares about, versus what impacted community members care about?

Assumptions behind outcome choice.

- What assumptions are we making, when deciding what the tool should optimize?
- How are we operationalizing goals for the tool, e.g., improving child 'safety'? What assumptions are we making?
- How do we bring providers to the table to decide on the use of outcomes?

Predictability of outcomes.

- Have we run any tests on historical data records, to check whether we get predictions on this outcome that actually make sense?
- How rare is the event we are trying to predict? If it is rare, how reliably do we think we can predict it?
- How does the inclusion of additional information (e.g., attributes) improve outcomes?

A.1.3 ***Empirical evaluations of algorithmic tool.***

Measuring improvement based on outcomes.

- Once the tool is deployed and in use, how can we evaluate how well it is working in the short-term? How can we evaluate how well it is working longer-term?
- What are some ways we might evaluate whether this tool is successful in improving the targeted outcomes?
- For evaluating worker-ADS decisions post-deployment: Do the decisions change by worker experience, worker demographic, or by supervisor?
- What performance measures do we plan to use to evaluate the tool?
- What performance measures have already been used in early analyses of historical data, prior to the deployment of the tool?
- Does this tool improve outcomes? How are we operationalizing "improve"?
- How does the use of the tool compare with the status quo? E.g., can we demonstrate the tool improves outcomes for the population of interest?
  - What is the "performance" and "fairness" of the existing baseline/status-quo decision process?
  - Is there someone with relevant domain expertise that could help explain anomalies or trends?
- Do we think there are tradeoffs between accuracy, fairness, explainability? If so, what are they?
- How are we measuring negative and positive impact on families?
- Is there someone with relevant domain expertise that could help explain anomalies or trends?
  - How well do you understand the domain application of the historical data used in evaluation?
  - Are there changes in policies and domain-specific practices in the historical data?
- Are there measured improvements resulting from the model's deployment?

- Are we using appropriate evaluation methods, e.g., synthetic controls, discontinuity analysis when cutoffs on risk exist.
- What outcome measures are we evaluating on? What can these measures tell us, and what can they not tell us?

Centering community needs.

- How can we effectively evaluate the tool from the perspective of impacted community members?
  - E.g., what does false positive, false negative mean for different impacted communities? How are we weighting false positives and false negatives, in a given use case, based on the relative costs of each type of error for the impacted stakeholders?

Worker perceptions.

- How might front-line workers respond to the tool? How can we better understand their underlying concerns and desires towards the tool?
- How do front-line workers perceive the algorithm? (e.g., do they consider it a top-down requirement or a useful tool)
- Do domain experts also believe the model 'makes sense', e.g., selection of important features?

A.2 Societal and Legal Considerations

The set of questions below are intended to support conversations around the following broader question: **Given the societal, ethical, and legal considerations and envisioned impacts associated with the use of AI tools for our stated goals (identified in Facet 1), is our proposed AI tool appropriate?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: AI practitioners, frontline workers, community members, legal experts.

A.2.1 ***Legal considerations around the use of algorithmic tool.***

- Do the people impacted by the tool have the power or ability to take legal recourse?
- Is there clarity around policies, e.g., whether algorithmic outcomes are included under 'public records'?
  - If someone asks for information around the tool, but there's no precedent, does the agency know what to do?
- Are you having conversations with the Department of Justice and attorneys, to make sure the new decision models you implement will follow existing policies, procedures, statutes, and rules?
  - Do you know which design decisions will be dictated by the law? For example: In the context of child maltreatment screening, if certain conditions are present in a case, then it is legally required to screen in for investigation.
- Can you inform existing policies, procedures, statutes, and rules to better meet the needs of new decision models?
- Do you need a new temporary rule to receive permission to use the model?
- How are you interpreting challenges to ambiguities in prior legal decisions around the use of the tool?
- What are challenges to interpreting legal documentation and guidelines?
  - How well can we interpret case-specific considerations in the context of legal documentation/guidelines (e.g., when there is a lot of grey in practice, but the law is written in black and white)?
    - E.g., in child maltreatment: "threat of harm" or "physical abuse" allegation type sounds black/white but there are various factors that make this grey. E.g., how hard did it hit them? Did it leave a mark? Action occurred but no impact from the action?

A.2.2 ***Ethical and fairness considerations around the use of algorithmic tool.***

Impacted Community Member Needs.

- Are there differences in the goals the agency versus community members think the tool should address? If so, what are they? If you are uncertain, what are your plans for understanding potential differences?
  - What are the envisioned harms and intended benefits from the tool that impact the community and the agency?
- Can we have impacted community's representatives or advocates at the table, to inform the design and use of the tool?
- How well are we engaging people closest to the problem and those impacted through the entire design, development, implementation, maintenance process?
- Are the outcomes intended for agency or community benefit?
- How well do we understand what outcomes the community wants to improve?
- Do we understand how impacted stakeholders perceive each decision? E.g., emotional valence, potential impacts, etc.
- To what extent are we optimizing the things the agency cares about versus what impacted community members care about?

Involving Impacted Communities.

- What are underlying assumptions that tool developers/researchers may have, regarding the soundness of the design decisions made in the tool?
- How can we set up external participation opportunities, to increase access?
  - E.g., avoiding scheduling during a 9-5pm period (to open involvement to those who want to be involved)
  - E.g., is it possible to involve groups that are not involved and paid by the agency, to get input and feedback?
  - Do we know who should be included? How can we build the right network of people to talk with?
- Who has a seat at the table, to decide how the tool impacts you?
- How are you engaging with people closest to the problem (e.g., frontline workers, community members, or others impacted by the decisions)?
- Have you communicated the limitations and historical context of the data, to community members?
- How well do we understand the costs, risks, and effort required of community members, if we invite them? E.g., many were directly harmed by decisions made by the agency.

- When do we start to engage impacted communities into discussions around the design or use of the tool?

Clarity of Ethics Goals and Definitions.

- Can we agree on a definition of fairness and equity in this context? What would it look like if the desired state is achieved?

Operationalization of Ethics Goals.

- Are fairness and equity definitions and operationalizations adequately context-specific? (For example, in the child welfare domain: children with similar profiles receive similar predictions irrespective of race?)
- Do we know how to appropriately operationalize our fairness formulation in the algorithm design?
- Can we mitigate biases in the model?
- How can we balance tradeoffs between false negatives and false positives?
- How well are we integrating domain-specific considerations into the design of the tool?
- Have we recognized and tried to adjust for implicit biases and discrimination inherent in these social systems that might get embedded into the algorithm?

Envisioning Potential Negative Impacts.

- Do we understand the negative impacts of the decision made across sensitive demographic groups?
- What are the externalities / long-run consequences of the decisions?

A.2.3 ***Social and historical context surrounding the use of algorithmic tool.***

- Have we recognized and tried to adjust for implicit biases and discrimination inherent in these social systems that might get embedded into the algorithm?
- How might we clearly communicate the limitations and historical context of the data to community members?
- Are you modeling historical, systemic patterns?

A.3 Data and Modeling Constraints

The set of questions below are intended to support conversations around the following broader question: **Given the availability and condition of existing data sources, and our intended modeling approach, is our proposed AI tool appropriate?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: AI practitioners.

A.3.1 ***Understanding data quality.***

- How does the data quality and trends compare with an 'ideal' state of the world?
  - What does our data look like, in terms of different demographic outcomes?
- Has the definition of the data changed over time? (E.g., in child welfare, has reunification always meant to reunify with the parent?)
- What data do we have access to?
  - Do we have the data/feature set to replicate the tool/analysis/and predictive accuracy of the existing tool?
- How well do we understand the meaning and value of the data that will be used to train an algorithm?
- How is the quality of this data?
  - How accurate is the data?
  - How recent is the data?
  - How relevant is the data?
  - Has the data been consistently collected?

A.3.2 ***Process of preparing data.***

- How are we preprocessing the data?
- Who should be involved in making decisions around whether to include or exclude certain data points or features? Do we have plans for involving those people?
- How do we address bias in the data?
- Do we have metrics for feature importance, that we could show relevant domain experts?
- How well do we understand the data collection process?
- Data leakage questions: Are we preventing oversampling of certain populations?
  - E.g., in child welfare: Are we pulling one child per report, and one report per child, to ensure there's no information leakage between training and test sets?

A.3.3 ***Model selection.***

- Is our model appropriate given the available data? Why or why not?

A.4 Organizational Governance Factors

The set of questions below are intended to support conversations around the following broader question: **Given our plans for ensuring longer-term technical maintenance and policy-oriented governance, do we have adequate post-deployment support for our proposed AI tool?** This stage would benefit from the expertise of the following stakeholders at the minimum, amongst others: Agency leaders, AI practitioners, frontline workers.

A.4.1 ***Long-run maintenance of algorithmic tool.***

Measuring changes in model performance over time.

- Do we expect there will be shifts in performance metrics over time? If so, why? What are our plans for identifying and mitigating those shifts?
- Do we expect that the data collection process will improve over time? What might this imply for how we maintain the tool? E.g., Is there a need for adjusting thresholds over time?

Mechanisms to identify long-run changes.

- Are we repeating feature engineering efforts over time?
  - Are we detecting how trends shift over time at the population level?
- Are there mechanisms in place that track whether certain data features have changed over the years?
- Do we have mechanisms to track longer-term outcomes over time, so that we can monitor for changes in model performance?
- Do we have the mechanisms to monitor whether the tool is having unintended consequences?

A.4.2 ***Organizational policies and resources around the use of algorithmic tool.***

Policies around worker interactions.

- Is there training for frontline workers who will be asked to use the tool? What evidence suggests that this training is adequate?
- How are frontline workers trained?
- Is it clear to workers what information the tool can access, and what information it cannot?
  - How is this communicated to workers?

Governance structures.

- Imagine that we could assemble the "ideal team" to monitor and govern the tool after it is deployed: What are the characteristics of this ideal team?
  - Who is the *actual* team that will monitor and govern the tool after it is deployed?
  - Given the gaps between the "ideal team" and the actual team we expect to have: What risks to post-deployment monitoring and governance can we anticipate? How might we mitigate these risks?
- Are there appropriate forms of governance, around the implementation?
  - Do those involved in governance have domain knowledge in the application context and have knowledge of the implementation process?
- Are there sufficient guardrails in place to ensure algorithms wouldn't get weaponized?
  - E.g., IRB-like programs and researchers at the same table, to minimize risk of weaponizing?

A.4.3 ***Internal political considerations around the use of algorithmic tool.***

- How well do we understand system administrators' and leadership's perspectives around the use of this tool?
- How well do staff and leadership understand 'why' the tool could bring value?
- Do system administrators and leadership perceive this tool positively?
- Do leadership support the future use of the tool?
  - Do we have backing at a leadership level? E.g., director, agency, governor, community partners?
- Is there sufficient buy-in from middle managers and executive support?
- Do we have mechanisms to address concerns that could come up during the ideation and design process?